FACT5: A Novel Benchmark and Pipeline for Nuanced Fact-Checking of Complex Statements Using Low-Resource Large Language Models

Shayan Chowdhury[1], Sunny Fang[2], Smaranda Muresan[2],

[1]Department of Computer Science, Columbia University, New York, NY, USA
[2]Department of Computer Science, Barnard College, Columbia University, New York, NY, USA

Abstract

Fact-checking complex statements is integral to combating misinformation, but traditional manual approaches are time-consuming, and recent automated approaches often oversimplify truthfulness into binary classifications and rely on resource-intensive models. This paper introduces two new developments: (i) FACT5, a curated dataset of 150 real-world statements with five ordinal classes of truthfulness, designed to capture the nuanced nature of factual accuracy and (ii) an end-to-end pipeline using large language models (LLMs) that decomposes statements into atomic claims, generates targeted questions, retrieves evidence from the web, and produces justified verdicts. Using FACT5, we evaluate our pipeline's performance using MISTRAL-7B-V0.3 and Google's GEMINI-1.5.FLASH. Our findings demonstrate significant improvements over baseline LLM performance, with MISTRAL-7B showing a 71.9% reduction in MSE for pass@3 evaluation. Notably, our pipeline achieves strong results using low-resource and/or open-source models that can run on consumer hardware, suggesting a path toward improving accessibility for AI-assisted fact-checking while maintaining high standards of accuracy. The FACT5 dataset, pipeline implementation, and evaluation framework is anonymized and provided at https://anonymous.4open.science/r/LLM-FactChecker/, and a demo of the pipeline can be found here https://fact5check.streamlit.app/.

1 Introduction

Traditionally, fact-checking relied on time-consuming and resource-intensive work by human experts (e.g., crowd-sourcing). With the widespread dissemination of mis/disinformation, coupled with growing capabilities in natural language understanding and generation of large language models (LLMs), recent work has been exploring how to leverage LLMs for automated fact-checking. Wang et al. (2024) proposed Factcheck-GPT, an approach that uses retrieval-augmented methods to detect and correct factual errors in natural language, including text generated by LLMs which can be prone to hallucinations. Gou et al. (2023) introduced the CRITIC framework for LLM self-correction, which extends output through iterations of verification and correction.

When handling claims that require reasoning, existing methods fall short in capturing the nuance of factuality with binary true/false labels (Vlachos and Riedel, 2014). Factcheck-GPT by (Wang et al., 2024) addresses this issue by breaking down complex claims into a series of atomic facts to verify separately, whereas Min et al. (2023) proposed more fine-grained metrics such as FACTSCORE, computing the percentage of atomic facts supported by reliable knowledge sources.

Existing LLMs-aided fact-checking methods have made significant breakthroughs, but it remains challenging to fact-check long-form statements that are decomposable, necessitates multiple resources, and usually requires multi-hop reasoning. To tackle these challenge, our work makes the following contributions: (i) curated test dataset—named **FACT5**: **F**actual **A**nalysis of **C**omplex **T**ruths (**5**-label)—of 150 statements with a five-class ordinal scale for truthfulness classification, (ii) a comprehensive end-to-end pipeline that supports ranked web-based search, multi-hop reasoning, question-answering, and five-way classification, and (iii) increased transparency and explainability with source citation and reasoning for each intermediary step.

2 Related Work

In recent years, automated fact-checking has gained traction in the journalistic process, from pioneers such as ClaimBuster (Hassan et al., 2017), to novel approaches such as one that leverages a frame-semantic parser (FSP) (Devasier et al., 2025). Guo

et al. (2022) comprehensively reviewed the state of fact-checking research as of 2022. A common theme emerges in recognizing that binary factuality falls short in capturing factual correctness in abstractive or complex (e.g., political) settings. Thus, research has developed multi-level typologies or turned to sources such as PolitiFact as potential datasets (Wang, 2017; Ma et al., 2023; Devasier et al., 2025), which we took note when designing our dataset for evaluation detailed in §3.1. Other benchmarks such as AVERITEC also moved past the binary and took web-evidence into account in claim verification (Schlichtkrull et al., 2024), but challenges remains as Pagnoni et al. (2021) notes how conventional metrics in natural language processing (NLP) such as METEOR score fall short in measuring factual correctness of generated reasoning. On top of veracity classification, quality evaluation of textual justifications represents an emerging direction for fact-checking frameworks (Russo et al., 2023). Our fact-checking pipeline, detailed in §3.2, is outlined in the following steps: (1) atomic claim generation, (2) question/query generation, (3) retrieval of relevant documents, (4) answer synthesis, (5) claim-wise classification, and (6) overall statement classification and justification. Similar approaches have been seen in past research (Wang et al., 2024; Rothermel et al., 2024; Wei et al., 2024). Other work has focused on investigating intermediary steps, such as atomic claim generation (Gunjal and Durrett, 2024; Wanner et al., 2024a) and found that conducting decomposition and decontextualization in one step yields optimal results (Wanner et al., 2024b), which we incorporated into our pipeline.

Furthermore, while recent approaches have shown promising results using LLMs, they often rely on resource-intensive architectures that require significant computational resources or expensive API calls. This creates barriers for smaller organizations and independent researchers who may lack access to such resources. Our work specifically addresses this limitation by demonstrating that comparable or superior performance can be achieved using lightweight, cost-effective models or ones that can run on consumer hardware.

3 Methods

3.1 Test Dataset

The need for a new dataset (FACT5) stems from several critical limitations in existing fact-checking datasets. Current datasets predominantly use binary true/false labels, which fail to capture the nuanced nature of factual correctness in complex statements. While AVERITEC represents an important benchmark in fact-checking research, our decision to evaluate primarily on FACT5 was motivated by several key factor, including label granularity and temporal relevance.

Additionally, concerns about data contamination and model memorization necessitate fresh data collection (Balloccu et al., 2024; Carlini et al., 2022). As LLMs may have been trained on existing fact-checking datasets, evaluating their true fact-checking capabilities requires testing on previously unseen claims. Our dataset's temporal range (January 2024 to January 2025) ensures the evaluation of model performance on genuinely novel information rather than memorized training data.

Sources	Verdicts				
	F	MF	HT	MT	T
PolitiFact	21	24	22	24	20
Snopes	4	0	1	1	4
WaPo	9	3	2	0	0
CNN	8	0	2	2	3
Total	42	27	27	27	27

Table 1: Summary of Dataset
F = FALSE, MF = MOSTLYFALSE, HT = HALFTRUE, MT = MOSTLYTRUE, T = TRUE

Our dataset, called **FACT5**, contains of 150 claims from the date range of January 10th, 2024 to January 31st, 2025. Our dataset exclusively draws from recognized fact-checking institutions, including PolitiFact[1], Snopes.com[2], The Washington Post's Fact Checker Section[3], and CNN's Facts First[4]. The distribution of sources is summarized in Table 1.

A key methodological decision was to prioritize sources that provide gold label evaluations alongside fact-check analyses, whose labels can be mapped to PolitiFact's *Truth-O-Meter* labels. This requirement was essential for our classification objective, though it necessarily narrowed the pool of eligible source material. *Truth-O-Meter* labels are provided in Table 5 in Appendix A.1. Mappings for the labels among data sources are in Appendix A.2, Table 6. A snippet of the dataset

[1] www.politifact.com
[2] www.snopes.com
[3] www.washingtonpost.com/politics/fact-checker
[4] www.cnn.com/specials/politics/fact-check-politics

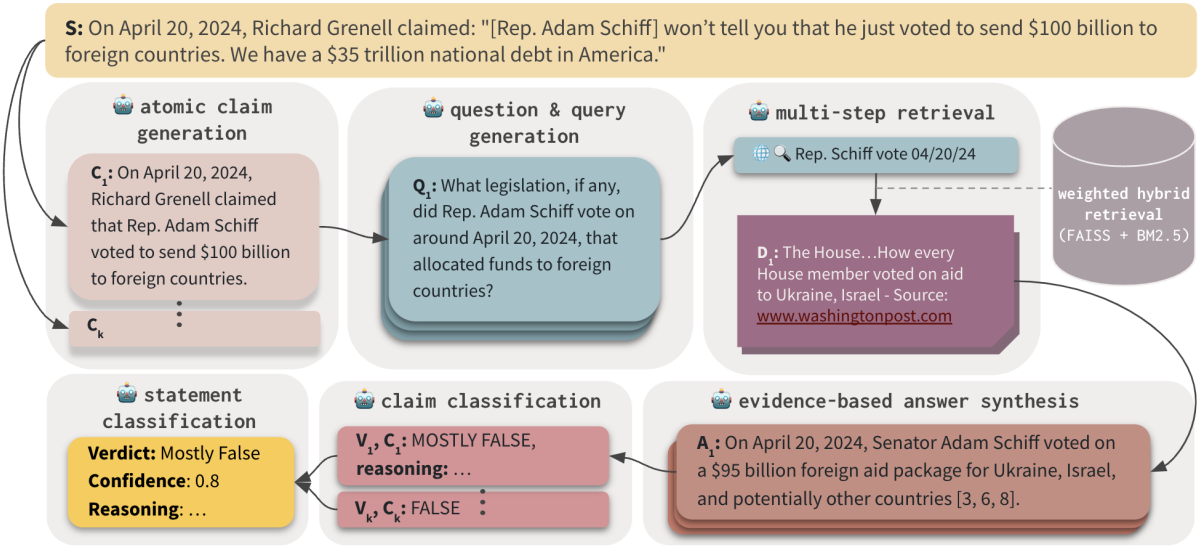

Figure 1: Overview of our proposed pipeline. Each gray box indicates a step detailed in §3.2

can be found in Appendix A.3.

Based on the *Truth-O-Meter*, the five classes used to classify a given statement are TRUE, MOSTLYTRUE, HALFTRUE, MOSTLYFALSE, and FALSE, representing an ordinal scale of factuality. The verdict UNVERIFIABLE is provided as an option for models to explicitly state when there is insufficient evidence to make a verdict. Statements labeled as FALSE has $n = 42$ while the others have $n = 27$ each since in real-life settings, it is more useful to assess a statement's falsehood.

3.2 Fact-Checking Pipeline

This section details each step of our enhanced fact-checking pipeline, visualized in Figure1.[5] While in an initial version of this paper, we did rigorous prompt engineering for each step, we now leverage the `DSPy` library (Khattab et al., 2022, 2023), a framework to optimize the outputs of language models through a declarative programming approach instead of manual prompt engineering. Each of these steps also leverages chain-of-thought (CoT) prompting to elicit improved reasoning capabilities (Wei et al., 2022).

Step 1: Atomic Claim Generation. Given a `statement` from the dataset in §3.1, the model is prompted to decompose and decontextualize the statements into **`atomic_claims`** in the form of a one-dimensional list of strings. Each claim should not rely on additional context to be understood and should focus on a single idea or concept (Barsalou, 1982; Wang et al., 2024).

Step 2: Question & Query Generation. For each claim in **`atomic_claims`**, the model is prompted to generate two key components: (1) questions that break down claim into verifiable sub-components and (2) search queries optimized for retrieving relevant evidence.

Step 3: Multi-Stage Retrieval of Relevant Documents. We implement a custom retrieval-augmented generation (RAG) system (Lewis et al., 2020) that involves fetching information from external sources (i.e. the internet) and a hybrid retrieval approach combining dense and sparse retrieval. For each claim, we iterated through the list of **`questions`** and conducted two sub-steps for each question:

Step 3a: Querying. Using the queries generated for each question from Step 2, we conduct web searches via API calls. We have implemented functionality to use both DuckDuckGo as well as Google Search via Serper's Search Engine Results Page (SERP) API[6], which returns a list of **`search_results`** that each includes the title, excerpt provided by the search engine, and metadata for the website.

Step 3b: Dense-Sparse Hybrid Retrieval. We first split the **`search_results`** retrieved from the web into chunks and processed them for retrieval

[5] For each, see Appendix B for criteria & `DSPy` signature

[6] https://serper.dev/

using a dual-index system as demonstrated in Wang et al. (2021). For dense retrieval, we utilize the `all-MiniLM-L6-v2` pre-trained sentence embedding model (Wang et al., 2020) using the Sentence-Transformers library to calculate vector representations of each text chunk, then store them in a vector database—in this case, Facebook AI Similarity Search (FAISS) (Douze et al., 2024) due to its efficiency and ease of implementation, but any other vector database such as ChromaDB could also be used. Simultaneously, for sparse retrieval, we implement `BM25`, the keyword text-retrieval algorithm using the `BM25Okapi` library for traditional lexical matching. To combine these two retrieval methods, we use a weighted combination ($\alpha * BM25 + (1 - \alpha) * FAISS$) to determine final document relevance, ensuring that the retrieved documents are both semantically and lexically similar to the query—similar to how web search engines work. We then retrieve 10 documents with the highest combined relevance scores to help answer each question in the following step.

Step 4: Evidence-Based Answer Synthesis. For each question, the pipeline synthesizes answers using the relevant evidence. Since each chunk of evidence retrieved retains metadata regarding the source, we can maintain provenance through explicit source attribution and inline citations. Furthermore, the pipeline also tracks the relevance score of each document to the question to help with the synthesis process to weigh the importance of each document in the final answer. This process is then repeated for every single question-answer pair for each claim.

Step 5: Claim-Wise Classification. With a list of question-answer pairs and evidence for each atomic claim, the claim is then evaluated for truthfulness along with a reasoning.

Step 6: Overall Statement Classification. Similar to the claim-wise classification step, the overall statement containing all the claims is then evaluated for truthfulness. The final verdict for the statement is determined by considering the atomic claims—and each of their question-answer pairs, verdicts, and confidence scores from step 5—inter-claim relationships, and the original statement.

Since the truthfulness reasoning of each claim contains information pertinent to determining the overall statement's truthfulness, we harness the reasoning capabilities of the model (Zhang and Gao, 2023). Adopting the same class labels from the claim-wise classification in the previous step, we finally classify the overall statement into one of the six classes.

3.3 Language Models Used & Technical Specifications

Models used for our research include GEMINI-1.5-FLASH and MISTRAL-7B-V0.3. MISTRAL-7B is an open-weight model that uitlizes Grouped Query Attention (GQA) and Sliding Window Attention (SWA) to improve performance and lower cost (Jiang et al., 2023). MISTRAL-7B-V0.3 is built upon its previous versions with a vocabulary of 32,768 tokens, enhancing the model's language understanding and generation capabilities (Jiang et al., 2024). Google's GEMINI-1.5-FLASH is designed for high-volume, cost-effective applications. It is online-distilled from GEMINI-1.5-PRO, a sparse mixture-of-expert (MoE) model; its number of parameters is not disclosed but can be reasonably estimated to be somewhere between 8B and 200B (Team et al., 2024).

We chose these two models specifically due to their cost-effectiveness and performance to maximize accessibility and ease of use: MISTRAL-7B is open-source and can be run locally on many consumer-grade hardware, while GEMINI-1.5-FLASH has a "free tier" with limited rate limits but is still a very powerful, versatile, and fast model. We ran our experiments on a MacBook Pro with an M1 Pro processor and 16GB of RAM using Ollama (Ollama, 2024) to leverage MISTRAL-7B for local inference, taking roughly 2 GPU hours for one pass through our entire FACT5 dataset. In total, we ran 3 passes through the dataset for each model, taking roughly 6 GPU hours for MISTRAL-7B and 2.5 GPU hours for GEMINI-1.5-FLASH. All models were run with a temperature of 0.3 and a maximum context length of 8192 tokens.

4 Evaluation

4.1 Ablation Studies

Pipeline aside (§3.2), we evaluated how well the LLM predicts the factuality label when only the statement itself is provided. The two main methods tested are as follows:

- **Baseline**: Only the `statement` is given to the model to generate a factuality label.

- **Pipeline**: After iterating through the proposed pipeline (§3.2), the statement, atomic claims, question-answer pairs, and claim assessment are given to the model to generate a factuality label.

If providing relevant information queried from the internet enhances the model's fact-checking capability, it would demonstrate the model's ability to effectively synthesize and reason over external knowledge sources, a desirable trait for reliable automated claim verification systems. The baseline condition provides a basis for comparison to see if the model can answer accurately without external information. Since not all language models have an explicit cutoff date, a fair baseline performance makes it challenging to know if the correct answer stems from memorization.

4.2 Evaluation Metrics

As mentioned in §3.1, our work treats fact-checking as an **ordinal multi-class classification** task. Our evaluation framework first mapped ordinal verdict classes to numerical values (TRUE = 5, MOSTLYTRUE = 4, HALFTRUE = 3, MOSTLYFALSE = 2, FALSE = 1). Crucially, although not in the gold dataset, the label UNVERIFIABLE is a possible output for the LLMs at the classification step, erring on the side of caution when there is insufficient evidence. For ordinality-based metrics, we excluded UNVERIFIABLE verdicts from calculations. This decision was motivated by two key factors: the inherent difficulty in quantifying the "distance" between an UNVERIFIABLE verdict and other categorical verdicts, and the fundamentally different nature of UNVERIFIABLE claims, which indicate insufficient evidence rather than a position on the ordinal truth spectrum.

Drawing lessons from Kulal et al. (2019), we employ the pass@k metric when evaluating model outputs. Under this paradigm, we prompted the model k times for each statement. For ablation study detailed in §4.1, we extracted labels for pass@1 and pass@3. Specifically, for pass@1 evaluation, we considered only the first prediction and excluded UNVERIFIABLE responses, whereas for pass@3 evaluation, we sorted predictions by their MSE and selected the best non-UNVERIFIABLE prediction if it exists.

The ordinal nature of classes calls for an evaluation metric that penalizes misclassifications that are "further" away from the gold label. For example, misclassifying FALSE as MOSTLYFALSE should be less penalized than misclassifying it as MOSTLYTRUE. Several studies have investigated the most appropriate way of handling ordinal classification (Cardoso and Sousa, 2011; Sakai, 2021; Amigó et al., 2020). Literature suggests that Mean Squared Error (MSE) remains to be a better metric when severity of errors weigh more (Gaudette and Japkowicz, 2009). MSE, thus, serves as our primary metric, with lower values indicating better performance.

Another evaluation metric measures the inter-rater agreement between expert fact-checkers (i.e., verdict from our dataset) and LLMs. Cohen's *quadratic weighted* κ is well-suited for ordinal multi-class classification (Cohen, 1968; Yilmaz and Demirhan, 2023). Similar to MSE, disagreements that are further weigh more with quadratic weights. The metric ranges from -1 to 1, with values closer to 1 indicating better agreement. We conducted listwise deletion (i.e., dropping statements if prediction is UNVERIFIABLE) as suggested in the findings of De Raadt et al. (2019).

Macro-average metrics remain crucial in evaluating multi-class classification performance although ordinality is not considered. Macro metrics and balanced accuracy consider the overall performance without taking into account class sizes, which is well-suited for our purpose since correctness regardless of class is crucial for fact-checking (Grandini et al., 2020). §5 discusses results in detail.

5 Results

We evaluated model architectures mentioned in §3.3 and §4.1 with metrics detailed in §4.2.

5.1 Comparative Analysis

We observe that both model implementations demonstrate the effectiveness of our pipeline, with Mistral showing larger relative improvements over its baseline (55.8% and 71.9% for pass@1 and pass@3) compared to GEMINI (43.7% and 67.1%). The pipeline consistently showed better coverage across both models, particularly with MISTRAL where it could make verifiable predictions in nearly all cases (147/150 for pass@3). For pass@3, our methodology of selecting the best non-UNVERIFIABLE prediction among the top three responses allowed both systems to improve their performance compared to pass@1, with the pipeline showing particularly strong gains.

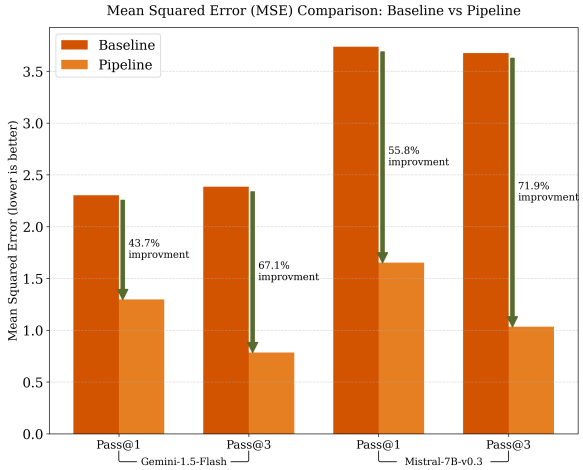

Figure 2: Comparison of Mean Squared Error (MSE), showcasing improvement of pipeline on multi-class ordinal truthfulness classification task, with sample size (out of 150) as follows (UNVERIFIABLES excluded):
GEMINI: pass@1: base: n=89, pipeline: n=101; pass@3: base: n=96, pipeline: n=121
MISTRAL: pass@1: base: n=72, pipeline: n=135; pass@3: base: n=89, pipeline: n=147

The stark difference in improvement percentages between pass@1 and pass@3 reveals an interesting characteristic of our pipeline. While both systems benefit from multiple prediction opportunities, our pipeline shows a more pronounced improvement with additional chances, suggesting that the pipeline maintains a more reliable ranking of alternative verdicts. Even when the first prediction isn't perfect, the correct verdict is more likely to appear in subsequent predictions, and the pipeline's uncertainty estimation is better calibrated, allowing it to generate meaningful alternative verdicts rather than just variations of the same prediction. In turn, the pipeline also enhances the model's ability to reduce UNVERIFIABLE predictions while maintaining or improving accuracy. This suggests that our structured approach helps models make more definitive verdicts without sacrificing reliability.

pass@k	Ablation	GEMINI-1.5-FLASH	MISTRAL-7B-V0.3
1	baseline	0.434	0.232
	pipeline	**0.681**	**0.516**
3	baseline	0.444	0.284
	pipeline	**0.810**	**0.702**

Table 2: Cohen's κ by ablation

GEMINI-1.5-FLASH. As shown in Figure 2, for our implementation with gemini-1.5-flash, our pipeline demonstrated substantial improvements over the baseline. In pass@1 evaluation, where we considered only the first prediction and excluded UNVERIFIABLE responses, the pipeline achieved an MSE of 1.3 compared to the baseline's 2.3, representing a 43.7% reduction in error. As for Cohen's quadratic weighted κ, the pipeline achieved 0.68 compared to the baseline's 0.43 on a [-1,1] scale, showing a 56.9% improvement. The pipeline also maintained better coverage, handling 101 out of 150 cases (with 49 UNVERIFIABLE predictions excluded) compared to the baseline's 89 cases (61 UNVERIFIABLE excluded).

The improvement was even more pronounced in pass@3 evaluation, where we sorted predictions by their MSE and selected the best non-UNVERIFIABLE prediction. Here, the pipeline achieved an MSE of 0.8 compared to the baseline's 2.4, marking a 67.1% reduction in error. Similarly, the pipeline outperformed in terms of Cohen's quadratic weighted κ, showcasing a 82.5% improvement with 0.81 compared to the baseline's 0.44. The pipeline successfully processed 121 cases (excluding 29 cases where all predictions were UNVERIFIABLE) while the baseline handled 96 cases (54 UNVERIFIABLE excluded).

MISTRAL-7B-V0.3. Figure 2 also demonstrates our implementation using Mistral-7B showing even stronger improvements. In pass@1 evaluation, the pipeline achieved an MSE of 1.6 compared to the baseline's 3.7, representing a 55.8% improvement. The pipeline achieved Cohen's quadratic weighted κ of 0.12 compared to the baseline's 0.06, demonstrating a 122.5% improvement. The pipeline also demonstrated significantly better coverage, processing 135 out of 150 cases (15 UNVERIFIABLE excluded) compared to the baseline's 72 cases (78 UNVERIFIABLE excluded).

For pass@3 evaluation, the pipeline achieved an MSE of 1.0 compared to the baseline's 3.7, showing a 71.9% reduction in error. Likewise, the pipeline achieved Cohen's quadratic weighted κ of 0.31 compared to the baseline's 0.09, demonstrating a 147.1% improvement. Moreover, pipeline maintained exceptional coverage with 147 cases (only 3 fully UNVERIFIABLE cases excluded) compared to the baseline's 89 cases (61 UNVERIFIABLE excluded).

5.2 Performance Analysis

On top of Cohen's quadratic weighted κ and MSE, we calculated additional classification metrics: bal-

Model	Ablation		Acc	Prec	Recall	F1
GEMINI	1	B	0.21	0.24	0.15	0.17
		P	**0.25**	**0.34**	**0.19**	**0.21**
	3	B	0.23	0.29	0.17	0.19
		P	**0.41**	**0.40**	**0.33**	**0.33**
MISTRAL	1	B	0.14	0.18	0.10	0.12
		P	**0.27**	**0.34**	**0.22**	**0.23**
	3	B	0.18	0.22	0.13	0.15
		P	**0.43**	**0.48**	**0.35**	**0.35**

Table 3: Balanced Accuracy and Macro Metrics (Precision, Recall, and F1-score) by LLMs and ablation: pass@k and B = baseline, P = pipeline (our method). Best performance in each group are in bold.

Gold Verdict	Top 3 Predictions		
FALSE (F)	*MF*	*F*	*HT*
	0.48	0.45	0.04
MOSTLYFALSE (MF)	*MF*	*HT*	*MT*
	0.52	0.22	0.15
HALFTRUE (HT)	*MT*	*MF*	*HT*
	0.40	0.33	0.22
MOSTLYTRUE (MT)	*MT*	*HT*	*MF*
	0.63	0.26	0.11
TRUE (T)	*MT*	*T*	*HT*
	0.44	0.22	0.22

Table 4: Top 3 predictions and respective proportions for pass@3 results from MISTRAL-7B-V0.3. Highlighted cells indicate exact match between gold and predicted.

anced accuracy, macro recall, precision, and F1 scores, as shown in Figure 3. While we excluded UNVERIFIABLE predictions from the MSE calculations to avoid distorting distance-based penalties—given that this class does not adhere to the ordinal structure—we included them in the macro calculations as it evaluates classification performance across all classes independently, without relying on ordinal relationships. This approach allows us to evaluate the models' performance across all possible outcomes, providing a comprehensive view of classification effectiveness.

The results demonstrate that our pipeline consistently outperforms the baseline across all metrics. Notably, the pass@1 pipeline configuration even surpasses the pass@3 baseline for both models, indicating the effectiveness of our approach even with limited passes.

Interestingly, MISTRAL-7B-V0.3 achieves the best overall performance among the tested models. This superior performance aligns with our earlier observation in the MSE analysis (Figure 2), where MISTRAL-7B-V0.3 showed a tendency to make fewer UNVERIFIABLE predictions. In the context of these classification metrics, this characteristic suggests that MISTRAL-7B-V0.3 may be more decisive in assigning specific truthfulness categories, potentially contributing to its improved performance across all classes.

To better understand our pipeline's performance on different verdicts, we analyzed the distribution of predictions for each true verdict class in our pass@3 evaluation. Table 4 shows the proportion of predictions made for each gold class, revealing several interesting patterns.

For FALSE claims, the pipeline shows strong discrimination between FALSE (45%) and MOSTLYFALSE (48%) verdicts, with minimal confusion with more positive verdicts. MOSTLYTRUE claims see the highest confidence predictions, with 63% of cases correctly identified. The pipeline shows some conservative tendency for TRUE claims, more frequently predicting MOSTLYTRUE (44%) than TRUE (22%). However, for HALFTRUE claims, the pipeline is more likely to predict MOSTLYTRUE (40%) than HALFTRUE (22%), and often got confused with MOSTLYFALSE as well—possibly due to the ambiguity of the verdict itself.

6 Discussion

The strong performance of our pipeline, particularly with the open-source MISTRAL-7B-V0.3 model, has significant implications for the democratization of AI assisted and potentially automated fact-checking. While larger proprietary models like GEMINI-FLASH-1.5 show impressive results, comparable performance achieved with MISTRAL-7B—a model that can run on consumer hardware—suggests that effective automated fact-checking need not be limited to organizations with extensive computational resources.

Furthermore, the strong pass@3 performance of our pipeline indicates that it may be effectively used in a semi-automated setting, where the system provides multiple ranked verdicts for human review. The verdict-wise analysis suggests that our system is particularly adept at identifying clearly false information. However, the conservative skew in more truthful predictions—where MOSTLYTRUE is preferred over TRUE, indicates a built-in caution against absolute certainty. This conservative tendency could be particularly valuable for fact-

checking organizations, as it helps avoid overconfident assertions while still maintaining clear discrimination between true and false information.

6.1 Practical Usage for Human Fact-Checkers

For independent fact-checkers and smaller organizations, the demonstrated effectiveness of our pipeline with MISTRAL-7B-V0.3 represents a compelling development. It suggests that robust fact-checking capabilities can be achieved without reliance on costly API calls to proprietary models or extensive computational infrastructure, significantly lowering the barrier to entry for assisted and/or automated fact-checking tools while upholding accuracy standards.

Furthrmore, a crucial feature of our pipeline is its flexibility in custom source document integration. While our current implementation primarily relies on web search and publicly available sources only, we designed our approach to also allow users to leverage their own proprietary databases and trusted sources—a capability particularly valuable for fact-checking organizations with extensive archives of verified information or specialized domain-specific knowledge bases. By enabling the use of custom source documents, the pipeline offers a sustainable pathway for scaling fact-checking operations without a proportional increase in costs—potentially making comprehensive fact-checking more feasible for a wider range of organizations.

6.2 Towards a Unified Benchmark

As mentioned earlier, the lack of a standardized benchmark modeling real-world fact-checking scenarios was a key motivation for developing FACT5. While recent work by Wang et al. (2025) and Tang et al. (2024) have made significant advances in automated fact-checking, their reliance on binary or ternary classification schemes (true/false/neutral) fails to capture nuance. Professional fact-checkers rarely deal in absolute truths or falsehoods, instead operating on a spectrum that accounts for partial accuracy, missing context, and degrees of misleading information. Our five-class ordinal classification approach—adapted from established fact-checking organizations' methodologies—addresses this limitation by providing finer-grained truthfulness assessment. This design choice makes direct performance comparisons with existing binary/ternary methods challenging, as mapping between different classification schemes would inherently lose critical nuance and potentially misrepresent the capabilities of each approach. For instance, a MOSTLYTRUE verdict in our method, indicating a statement that is accurate but requires clarification, would be forced into either TRUE or FALSE in a binary scheme—losing important context about the nature of any inaccuracies.

By introducing the FACT5 benchmark and our own method, we hope we can contribute to the development of more nuanced evaluation frameworks that move beyond binary classification. Our five-class ordinal scale and emphasis on complex, multi-hop reasoning requirements could serve as a foundation for future benchmarks that better capture the complexity of real-world fact-checking scenarios. The dataset's temporal recency and careful source attribution also address important considerations around data contamination and evaluation validity that should be standard features of fact-checking benchmarks moving forward.

6.3 Statement Metadata

Inspired by the work of Wang (2017), future research could investigate the effect of statement metadata on model performance. Our proposed pipeline has an input in the form "*On* `statement_date`, `statement_originator` *claimed,* '`statement`'." Current fact-checking research has not yet standardized an input, which warrants future researchers to conduct a systematic review of the impact of metadata on fact-checking tasks. For example, knowing the scenario (e.g., State of the Union Speech) during which the statement was made may help contextualize statements. Nonetheless, our pipeline attempts to model the need of fact-checking on the fly, such as live during a presidential debate, with minimal information available.

Future research directions should address the aforementioned dimensions to improve the accuracy, efficiency, and robustness in assessing the factuality of complex statements, as adaptive and scalable solutions are needed to combat misinformation and promote informed public discourse (Barman et al., 2024; Wirtschafter, 2024).

Limitations

While our research advances the automated fact-checking of complex statements using large language models, some limitations need to be carefully considered.

Dataset Limitations

Our current evaluation relies on our own curated FACT5 dataset of 150 statements, which represents a relatively small sample size compared to other NLP benchmarks. Though our pipeline shows promising results on the FACT5 dataset, its performance on a broader range of statement types and domains remains to be fully validated. The current evaluation, while thorough, may not capture all edge cases or statement complexities that could arise in real-world fact-checking scenarios.

To mitigate these limitations, we focused on high-quality, professionally fact-checked statements, ensured balanced representation across truthfulness categories, selected temporally relevant statements to test model performance on current claims, and incorporated multiple evaluation metrics for a comprehensive performance analysis. Future work should focus on expanding the dataset while maintaining these quality standards, potentially through collaboration with professional fact-checking organizations to access larger pools of verified claims.

The FACT5 dataset's recency (2024-2025) presents both advantages and limitations. While it allows testing of models' capabilities on current events, it also means that the dataset might become less relevant over time as the context of these statements changes, or if the training cutoff date for language models get extended to incorporate more recent web data. This temporal dependency could affect the long-term utility of both the dataset and the evaluation metrics derived from it.

Computational Resource Requirements

While our pipeline demonstrates strong performance with MISTRAL-7B-V0.3 on consumer hardware, the full pipeline including dense-sparse hybrid retrieval and multiple passes still requires significant computational resources. The time and resource requirements for processing multiple statements in parallel or handling high-volume fact-checking scenarios need to be carefully considered. This could pose challenges for real-time fact-checking applications, such as during live debates or breaking news situations.

Pipeline Robustness Concerns

The sequential nature of our pipeline means that errors in early stages (e.g., atomic claim decomposition or question generation) can propagate through the system and affect final verdicts. While our results show strong overall performance, the interdependence of pipeline components could make the system vulnerable to cascading failures, particularly when dealing with especially complex or nuanced statements.

For example, retrieval of top-k results from search engines serves as a fundamental component of our fact-checking pipeline. However, the inherent challenge lies in the lack of a definitive method to ensure the accuracy of the retrieved information. Despite prioritizing reliable sources and performing rigorous post-processing, the inherent accuracy of the information obtained cannot be guaranteed.

Need for Expert Human Evaluation

While we conducted a preliminary evaluation with participants lacking specialized knowledge in fact-checking, the results demonstrated limited value. The overwhelming agreement between untrained participants and model outputs suggests that this method may not provide sufficiently discriminative or insightful feedback for our purposes. Consequently, we have chosen to focus our analysis on more informative metrics detailed in §4.2, which are better suited to assess the performance of our five-way classification task for truthfulness. Nonetheless, as noted in Russo et al. (2023), evaluating the quality of model reasoning is also crucial beyond examining the correctness of the model outputs. Therefore, for future work, expert evaluators in journalistic fields or work as fact-checkers would be necessary for conducting robust human evaluation.

Challenges in Data Leakage

The absence of publicly accessible training data restricts our ability to explore the phenomenon of information memorization by LLMs for fact-checking purposes. Despite efforts to mitigate bias by blacklisting certain sources like PolitiFact, the ubiquity of its work across online content poses a challenge. Even if PolitiFact itself is excluded from the training data, its findings may still indirectly influence other sources used during the retrieval process, potentially impacting the reliability of the fact-checking outcomes.

Political Biases and Logical Fallacies

Previous work has exhibited that political leanings can be embedded into LLMs (Feng et al., 2023). Due to the nature of our research, it is possible that

LLMs exhibited political biases when determining the factuality of a statement, which could diverge from the nonpartisan nature of fact-checking tasks. A closer look into the results is needed to verify the presence of potential political biases in judging the factuality of statements. It is also important to note that although our model is primarily designed to fact-check complex statements, it is not yet equipped to identify common fallacies that are often deployed in political speeches (e.g., red herring and straw man fallacies).

Ethical and Privacy Risk Considerations

While our pipeline demonstrates promising capabilities for assisted fact-checking, the deployment of our system could inadvertently contribute to reduced trust in legitimate news when the model makes incorrect classifications. Furthermore, concerns around copyrighted material and privacy may also arise from our pipeline's reliance on web-based information retrieval and processing of potentially sensitive data.

To mitigate these risks, we recommend implementing human oversight in the verification pipeline (especially for high-stakes or sensitive claims), stricter data quality and minimization practices in the retrieval process, and establishing clear protocols for handling copyrighted materials retrieved from the web—including implementing similarity detection mechanisms to flag potential copyright conflicts or collaborating with news organizations to come to fair use agreements. Future work to minimize these risks should focus on developing more robust privacy-preserving techniques and establishing clear guidelines for responsible deployment of automated fact-checking systems.

References

Enrique Amigó, Julio Gonzalo, Stefano Mizzaro, and Jorge Carrillo-de Albornoz. 2020. An effectiveness metric for ordinal classification: Formal properties and experimental results. *arXiv preprint arXiv:2006.01245*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, page 100545.

Lawrence W Barsalou. 1982. Context-independent and context-dependent information in concepts. *Memory & cognition*, 10(1):82–93.

Jaime S Cardoso and Ricardo Sousa. 2011. Measuring the performance of ordinal classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08):1173–1195.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Alexandra De Raadt, Matthijs J Warrens, Roel J Bosker, and Henk AL Kiers. 2019. Kappa coefficients for missing data. *Educational and psychological measurement*, 79(3):558–576.

Jacob Devasier, Rishabh Mediratta, Akshith Putta, and Chengkai Li. 2025. Automatic fact-checking with frame-semantics. *arXiv preprint arXiv:2501.13288*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Lisa Gaudette and Nathalie Japkowicz. 2009. Evaluation methods for ordinal classification. In *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*, pages 207–210. Springer.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. 2020. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Anisha Gunjal and Greg Durrett. 2024. Molecular facts: Desiderata for decontextualization in llm fact verification. *arXiv preprint arXiv:2406.20079*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Albert Jiang, Alexandre Sablayrolles, Alexis Tacnet, Antoine Roux, Arthur Mensch, Audrey Herblin-Stoop, Baptiste Bout, Baudouin de Monicault, Blanche Savary, Bam4d, Caroline Feldman, Devendra Singh Chaplot, Diego de las Casas, Eleonore Arcelin, Emma Bou Hanna, Etienne Metzger, Gianna Lengyel,

Guillaume Bour, Guillaume Lample, Harizo Rajaona, Jean-Malo Delignon, Jia Li, Justus Murke, Louis Martin, Louis Ternon, Lucile Saulnier, Lélio Renard Lavaud, Margaret Jennings, Marie Pellat, Marie Torelli, Marie-Anne Lachaux, Nicolas Schuhl, Patrick von Platen, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Thibaut Lavril, Timothée Lacroix, Théophile Gervet, Thomas Wang, Valera Nemychnikova, William El Sayed, and William Marshall. 2024. Mistral-7b-v0.3.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*.

O. Khattab, K. Santhanam, X. L. Li, D. Hall, P. Liang, C. Potts, and M. Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, and C. Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S Liang. 2019. Spoc: Search-based pseudocode to code. *Advances in Neural Information Processing Systems*, 32.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. 2023. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Ollama. 2024. Ollama. `https://github.com/ollama/ollama`.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. *arXiv preprint arXiv:2104.13346*.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. Infact: A strong baseline for automated fact-checking. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 108–112.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Tetsuya Sakai. 2021. Evaluating evaluation measures for ordinal classification and ordinal quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2759–2769.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '21, page 317–324, New York, NY, USA. Association for Computing Machinery.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Yuxia Wang, Revanth Gangi Reddy, Zain Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi N Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2025. Openfactcheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and llms. In *Proceedings of*

the 31st International Conference on Computational Linguistics, pages 11399–11421.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024a. A closer look at claim decomposition. *arXiv preprint arXiv:2403.11903*.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024b. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *arXiv preprint arXiv:2412.13175*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Valerie Wirtschafter. 2024. The impact of generative ai in a global election year.

Ayfer Ezgi Yilmaz and Haydar Demirhan. 2023. Weighted kappa measures for ordinal multi-class classification performance. *Applied Soft Computing*, 134:110020.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*.

A FACT5 Pilot Dataset

A.1 Explanation of Ratings

Label	Description
True	The statement is accurate and there's nothing significant missing.
Mostly True	The statement is accurate but needs clarification or additional information.
Half True	The statement is partially accurate but leaves out important details or takes things out of context.
Mostly False	The statement contains an element of truth but ignores critical facts that would give a different impression.
False	The statement is not accurate.
Pants on Fire	The statement is not accurate and makes a ridiculous claim.

Table 5: Truth-O-Meter rating used by PolitiFact

A.2 Mapping of Fact-checking Metrics

Numeric Label	Our Mapping	Credible Sources		
		PolitiFact	The Fact Checker (WaPo)	Snopes
0	Unverifiable	N/A	No Verdict	Unproven / Unfouned
1	False	Pants on Fire / False	Four Pinocchios	False
2	Mostly False	Mostly False	Three Pinocchios	Mostly False
3	Half True	Half True	Two Pinocchios	Mixture
4	Mostly True	Mostly True	One Pinocchio	True
5	True	True	The Geppetto Checkmark	True

Table 6: Mappings of Fact-checking Metrics

A.3 Snippet of the Pilot Dataset

verdict	statement_originator	statement	statement_date	factchecker
FALSE	Donald Trump	"This year, the typical family's tax bill is thousands of dollars lower because of the Trump Tax Cuts..."	4/17/2024	factcheck.org/2024/04/trumps-unfounded-colossal-tax-hike-warning/
MOSTLY TRUE	Joe Biden	"100 million Americans can no longer be denied health insurance because of a preexisting condition..."	4/11/2024	factcheck.org/2024/04/familiar-claims-in-a-familiar-presidential-race/
MOSTLY TRUE	Ron DeSantis	"She spent 100% of her money attacking me..."	1/26/2024	cnn.com/2024/01/16/politics/fact-check-cnn-desantis-town-hall-new-hampshire/index.html

Table 7: Selected Rows and Columns from Pilot Dataset

B Pipeline Details

B.1 Atomic Claim Generation

DSPy signature for claim extraction, which consists of the criteria for each claim.

```
"""Extract specific claims from the given statement.
1. Split the statement into multiple claims, but if the statement is atomic (has one main claim), keep it as is.
2. If context is included (e.g., time, location, source/speaker who made the statement, etc.), include the context in each claim to help verify it. Do not make up a context if it is not present in the text.
3. Consider the source (e.g. name of the speaker, organization, etc.) and date of the statement if given in the context, and include them in each claim.
4. Each claim should be independent of each other and not refer to other claims.
5. Always extract claims regardless of the content """
```

Output field

```
"""JSON object containing:
{
"claims": [
{
"text": string, }
]
}"""
```

B.2 Question Generation

DSPy signature for question generation

```
"""Break down the given claim derived from the original statement to generate independent questions and search queries to answer it. Be as specific and concise as possible, try to minimize the number of questions and search queries while still being comprehensive to verify the claim."""
```

Output field

```
"""JSON object containing: {
"questions": [
{
"question": string, # question text (e.g. "What was the GDP growth rate during the Trump administration?")
"search_queries": [string], # independent search queries used to answer the question, try to be as specific as possible and avoid redundancy, 1-2 queries is ideal
}
]}"""
```

B.3 Answer Synthesis

DSPy signature for answer synthesis

```
"""Synthesize an answer based on retrieved documents with inline citations."""
```

Output field

```
"""JSON object containing:
{
"text": string, # answer with inline citations where the number in the brackets
is the index of the citation in the citations list (e.g., "The wage gap was
shrinking [1]")
"citations": [{ # list of citations
"snippet": string, # exact quote from source
"source_url": string,
"source_title": string,
"relevance_score": float
}]
}"""
```